JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-2 EXAMINATION (APRIL 2022)

B-Tech (6th SEM)

Course Code:18B1WCI635 Max. Marks: 25

Course Name: DATA MINING & DATA WAREHOUSING

Course Credit: 2 Max. Time: 1 Hour 30 Min

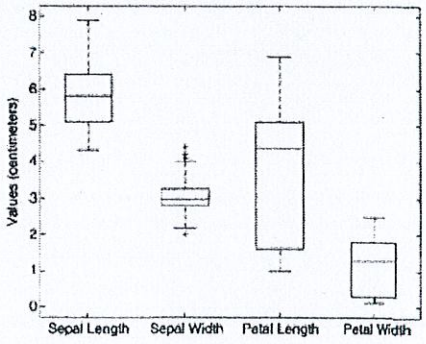
Note: All questions are compulsory

Q.1 [CO 2] Explain the following terminology in relation to performance evaluation of classification algorithm with example: (5)

i. Confusion Matrix
ii. TP, FN, FP, TN
iii. Accuracy
iv. Precision
v. Recall

Q.2 [CO 1, 2] (a) Discuss the advantages and disadvantages of using sampling to reduce the number of data objects that need to be displayed. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not? (2)

[CO 2, 3] (b) Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in illustration? (3)



Q.3 [CO 2] Generate Association Rule for Transactional data illustrated below by applying Apriori Algorithm? (5)

Transactional Data for an *AllElectronics* Branch

TID	*List of Item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Q.4 [CO 3] (a) Explain Entropy and Gini in terms of decision tree classification. Also provide comparison between these two using a diagram. (3)

[CO 2, 3] (b) Consider the training examples shown in Table for a binary classification problem. (7)

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

i. What is the entropy of this collection of training examples with respect to the positive class?

ii. What are the information gains of a1 and a2 relative to these training examples?

iii. For a3, which is a continuous attribute, compute the information gain for every possible split.

iv. What is the best split (among a1, a2 and a3) according to the information gain?

v. What is the best split (between a1 and a2) according to the classification error rate?

vi. What is the best split (between a1 and a2) according to the Gini index?